

# tym: Typed Matlab

Hamid A. Toussi

Department of Mathematics & Computer Science

University of Sistan and Baluchestan

Zahedan, Iran

e-mail: hamid2c@gmail.com

November 16, 2011

## 1 Introduction

Many scientists and engineers use Octave or MATLAB as their preferred programming language. However, dynamic nature of these languages can lead to slower running-time of programs written in these languages compared to programs written in languages which are not as dynamic, like C, C++ and Fortran.

Two dynamic features that contributes to performance issues in major ways are:

1. Dynamic typing: Types of variables can change during run-time and generally they are not known before run-time.
2. Dynamic resizing: Size of arrays and matrices can change during run-time. Pre-allocation of arrays is not mandatory in Octave/MATLAB and whenever a value is assigned to a location that is not within the range of the array indexes, the array is resized to store the new value.

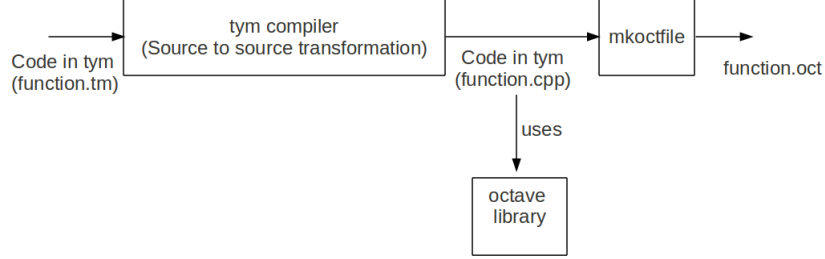
For example, consider the following function in Octave/MATLAB:

```
function z = mmt(x, y)
    z=x*y;
end
```

Type of variables  $x$  and  $y$  can be any allowable type. If we are supposed to compile this function statically, we may have to choose the widest possible type for  $x$  and  $y$ . This type is an array of complex numbers. The result would be very inefficient code when parameters are of narrower type (e.g. Integers or even arrays of integers). In these cases they are actually wrapped as arrays of complex numbers. One way to tackle this problem is to use JIT compilers to compile the program or choose the best previously compiled version at run-time. However, this requires run-time overhead and can be really non-trivial. We would like to follow another path in this paper.

We have developed a new language which is similar to Octave/MATLAB in many ways but has a less dynamic nature. In particular, variables must be declared (with their type) before they are defined or used and arrays must be allocated explicitly. We have called this language, tym (Typed MATLAB).

Figure 1: tym compiler architecture



Programs written in tym are translated into C++, the generated C++ program uses Octave library and can be called from the Octave [1] interpreter. To do this, an oct-file should be created by using mkoctfile.

Cython [4] also takes a similar approach to us but it is based on Python which might not be as common as Octave/MATLAB in scientific and engineering communities. OMPC [3] is another compiler that translates MATLAB to Python. We have used some of their routines and their grammatical rules in our code.

## 2 Overview

You can see different components that are necessary to translate a program in tym to its equivalent module in C++ in Figure 1.

Currently, a program should be written as a function in tym. Later, this function is transformed to a C++ module and is compiled to an oct-file. The resulted oct-file can be called from the octave interpreter as a function. This has the advantage that the user can change or write parts of her MATLAB program that is computation-intensive as a function in tym then call it from Octave so she has access to all packages in octave-forge (Neural networks, Image processing, Signal processing, ...) while she programs in tym.

Resulted C++ module uses Octave library classes and routines to manipulate matrices and other objects in a way that is compatible with Octave. However, since Octave library does not depend on the Octave interpreter and any C++ program can use it independent of the rest of the Octave, it is also possible to convert a program which is written in tym to an executable that can be run without relying on the octave interpreter.

To implement our tym to C++ compiler, we use PLY [2] which is a Python implementation of compiler construction tools lex and yacc. PLY supports LALR(1) parsing. All grammatical rules should be written in a python module (tymply.py in our compiler). There is a function for every tym programming construct in this module. The grammatical rule which corresponds to a construct is defined in the function's document string. In the body of the function is the action code which is done upon parsing the construct.

PLY generates a parser based on the file tymply.py. Since the generated parser is a bottom up LALR(1) parser, we can assume that it does a post-order traversal on Abstract Syntax Tree (AST) and performs some action code upon visiting each AST node. Every AST node is an instance of a tym construct in input

text. In our compiler, action codes are responsible for transforming the input program in tym to its equivalent form in C++.

AST is an abstract form of parse-tree which is constructed during parsing the input text.

### 3 Dynamic Linking and Octave library

Octave can be dynamically linked with functions which are written in C/C++. Upon linking any of the linked functions can be called from Octave interpreter. Writing Octave compatible functions requires either following the C-Mex interface or using Octave library to manipulate matrices and other objects you would like to pass to Octave. This is necessary to comply to Octave's memory representation of different objects (like matrices). Octave itself uses the Octave library to manipulate matrices and other objects so using it imposes no additional overhead.

Using C-Mex interface causes the overhead of converting input parameters of C-Mex routines to its equivalent Octave representation. However, it has the advantage that the functions that uses C-Mex interface can be also linked to MATLAB proprietary software.

We have used Octave library so the remainder of this Section is devoted to it and how a C++ function can be linked with Octave.

Consider the example in Figure 2. Using `#include <octave/oct.h>` is necessary so you have access to the required interface to Octava library.

`DEFUN_DLD` is a macro that defines the entry point to the function. It has four arguments, name of the function as appear in Octave, list of input parameters to the function, number of output parameters and function's document string. The function should always return an object of type `octave_value_list`. Input parameters are also passed as a `octave_value_list` object (`args`). Input parameters should be extracted from `args` based on their type. Octave passes the input parameters to the function as arrays. However, here, only two integers should be read which are stored in variables `x` and `y` respectively so the first element (indexed by 0) of each input array has been read. Then global variable `error_state` is checked to make sure that input parameters are of expected type. Finally, two values (`a` and `b`) that are supposed to be returned are stored into `retval` which is an `octave_value_list`. Note that operator `()` has been overloaded for `octave_value_list`.

In order to call this function from Octave, we have to save it as a file that has the same name as the function (`sumsub.cpp` int this example) and run `mkoctfile sumsub.cpp` command. Now we can call this function like any other function from octave interpreter.

The base of all matrices and arrays in octave library is `Array` class. All elements of an array is stored linearly in memory (see `data` in Figure 3). `len` member keeps the number of elements stored in the array. Dimensions and shape of the array can be achieved by referring to `dimensions` vector. Addresses required for doing lookup or assignment operation is calculated based on this vector. Calculated address is an integer that acts as an index for `data` member. The array can be easily reshaped just by making changes to `dimensions` vector. In fact, every `Array` object has a pointer to a `ArrayRep` object (`rep`) that contains members like `data` and `len`. This object also keep the number of

Figure 2: A simple function in C++ that can be linked with Octave

```
#include <octave/oct.h>
#include <iostream>
#include <cstdlib>
// File: sumsub.cpp
DEFUN_DLD (sumsub, args, nargout, "do summation and subtraction") {
    octave_value_list retval;
    if ((args.length()) != 2) {
        std::cout<<"invalid number of input params\n";
        return retval;
    }

    int x=args(0).int32_array_value()(0);
    int y=args(1).int32_array_value()(0);
    if (error_state) {
        std::cout<<"invalid type of input parameters\n";
        return retval;
    }
    int a = x + y;
    int b = x - y;
    retval(0) = a;
    retval(1) = b;
    return retval;
}
```

Array objects that share it in `count`. Every Array increments `count` of its `rep` on construction and decrements it on destruction. In `Array`'s destructor it is checked that whether `rep->count` has reached zero and free `rep` when it is so. `Array` is a parametrized type which uses parameter `T` to refer to type of the elements stored in the array. Array of different types can be instantiated by passing different types as `T`.

`Array` class has methods and other data members which are not shown in Figure 3. Many of its methods like `resize` and `reshape` should be familiar to any Octave/MATLAB user. Array lookup and assignment is done by calling either the methods `xelem` or `checkelem`. The former does not perform bound checking and the latter does.

Figure 3: Definition of Array class in Octave library

```
template <class T> class Array
{
    class ArrayRep {
        T *data;
        octave_idx_type len;
        int count; // reference count
        ...
    }
    dim_vector dimensions;
    ArrayRep *rep;
    ...
}
```

In Octave library `idx_vector` type is used to represent single indices, range slices and colons. Array slicing is done through two family of methods:

- **index** methods: Whenever, a sliced array is used in an expression, one of these methods will be called. For example, `a(1:4, 2:6)` can be implemented as `a.index(idx_vector(0, 4), idx_vector(1, 6))`. Note that indexing in Octave library is zero based and upper bounds in slices are exclusive.
- **assign** methods: Indexed assignment to arrays is done by using one of these methods. For example, `a(1:3, :)=b` can be implemented as `a.assign(idx_vector(0, 3), idx_vector::colon, b)`.

## 4 Types and Symbol Table Management

We have used a stack of symbol tables to implement nested scoping. These symbol tables are necessary to keep track of every identifier's type in input program. Other information like the line number where a variable is declared and the line number where it is defined are also kept to do various checks like whether a variable that is used, is declared and defined before its use.

Type information is used for various purposes including resolving ambiguity in the grammar in certain cases and very limited type inference and type checking.

## 5 Programming and experimenting with tym

Currently, you can write a function in tym, translate it to C++ and call it from Octave.

In contrast to MATLAB/Octave, variables must be declared before they are used. As I write this paper, there are only five types for variables in tym, namely `int`, `real`, `intArray` and `realArray`. Types `real` and `float` correspond to `double` and `float` types in C++. In Octave library, arrays of `double` (i.e `Array<double>`) are used to represent arrays of floating point numbers so only `realArray` is available in tym to avoid any extra conversion and copying. Hopefully, support for complex numbers and arrays will be added in the future.

There are also some directives in tym language that tells the tym compiler whether to generate code to do array bound checking, initialization of variables and similar stuff. Whenever they come in the tym program they enable/disable the desired feature from the line afterward.

As of this writing three directives are available. `$ 'zero_based_arrays'` tells tym compiler that matrices are zero based. `$ 'no_init_vars'` tells the compiler not to generate code for initializing variables and `$ 'no_check_ranges'` make the compiler generate code that does not do bound checking. All these directives are off by default. That is, when no directive has been presented in input program, the compiler generates code for one-based matrices, initializing variables and bound checking which is more consistent with Octave/MATLAB behavior. However, using any of these directive would have positive effect on performance of the generated C++ program, specially `$ 'no_check_ranges'`.

As an example consider the program in Figure 4. This program is a function in tym that multiplies two arrays of type `real`. All the mentioned directives are used to make it as efficient as possible.

Figure 4: A function in tym that multiplies its parameters

```
$ 'zero_based_arrays'
$ 'no_init_vars'
$ 'no_check_ranges'
% File: mymult.tym
function intArray z = mymult(realArray x, realArray y)
    int d1x = rows(x)
    int d2x = columns(x)
    int d1y = rows(y)
    int d2y = columns(y)

    if (d2x ~= d1y)
        error('incompatible dimensions')
    end

    createArray(z, d1x, d2y)

    int i
    int j
    int k
    for i=0:d1x-1
        for j=0:d2y-1
            z(i, j) = 0
            for k=0:d1y-1
                z(i, j) = z(i, j) + x(i, k)*y(k, j)
            end
        end
    end
end
end
```

Figure 5: Translated version of mymult

```
#include <octave/oct.h>
#include <iostream>
#include <cstdlib>
DEFUN_DLD (mymult, args, nargout, "") {
    octave_value_list retval;

    NDAarray x=args(0).array_value();
    NDAarray y=args(1).array_value();
    int d1x = x.rows();
    int d2x = x.columns();
    int d1y = y.rows();
    int d2y = y.columns();
    if ((d2x != d1y)) {
        std::cout<<"error"<<"incompatible dimensions"<<"\n";return retval;
    }
    int32NDAarray z(dim_vector( d1x, d2y));
    int i;
    int j;
    int k;
    for (i = (0); i <= (d1x - 1); i += (1)) {
        for (j = (0); j <= (d2y - 1); j += (1)) {
            z.xelem(i, j) = 0;
            for (k = (0); k <= (d1y - 1); k += (1)) {
                z.xelem(i, j) =
                    z.xelem(i, j) + x.xelem(i, k) * y.xelem(k, j);
            }
        }
    }
    retval(0) = z;
    return retval;
}
```

The tym compiler would translate the function in Figure 4 into the C++ code which is shown in Figure 5. To call this function from Octave you have to make a dynamically loadable Octave module out of it. To do this you can execute `mkoctfile mymult.cpp` in a terminal. The mentioned C++ program can be generated by invoking `python tymc.py mymult.tm`.

As an example for slicing see the program in Figure 6 which would be translated to the C++ code shown in Figure 7 by the tym compiler.

You can also try tymc interactively:

Type `python` and then `import tymply as t` in a terminal. Now you can translate a tym statement to C++ and see the result instantly. Just type `t.yacc.parse("tym_statement")` where `tym_statement` could be any valid tym statement, and press Enter.

## 6 Evaluation

We have done a limited evaluation using different versions of array multiplication. Results of this evaluation is shown in Table 1.

Three versions of multiplication have been implemented in tym. `mult-real` is the implementation which is shown in Figure 4. `mult-int` uses `intArray`

Figure 6: A function that add two array slices in tym

```
$ 'no_check_ranges'
% File: addslice.tm
function intArray z = addslice(intArray x, intArray y)
    if (rows(x) < 3 || columns(x) < 3 || rows(y) < 3 || columns(y) < 3)
        error('Matrices should be of size at least 3x3')
    end
    createArray(z, 3, 3)
    z = x(1:2, 1:2) + y(2:3, 2:3)
end
```

Figure 7: Translated version of addslice

```
#include <octave/oct.h>
#include <iostream>
#include <cstdlib>
DEFUN_DLD (addslice, args, nargout, "") {
    octave_value_list retval;

    int32NDArray x=args(0).int32_array_value();
    int32NDArray y=args(1).int32_array_value();
    if ((x.rows() < 3 || x.columns() < 3 || y.rows() < 3 || y.columns() < 3)) {
        error("Matrices should be of size at least 3x3");
        return retval;
    }
    int32NDArray z(dim_vector( 3, 3));
    z = ((int32NDArray)x.index(idx_vector(1-1, 2-1+1, 1), idx_vector(1-1, 2-1+1, 1)))
        + ((int32NDArray)y.index(idx_vector(2-1, 3-1+1, 1), idx_vector(2-1, 3-1+1, 1)));
    retval(0) = z;
    return retval;
}
```



Table 1: Running times of different versions of multiplication.

benchmark	$100 \times 100$	$300 \times 300$
mult-int	0.00457	0.1063
mult-int-check	0.0622	1.579
mult-real	0.002945	0.06881
mult-octave	19.68	155.3

instead of `realArray`. Version `mult-int-check` is similar to `intArray` except that it does not contain directives for not doing bound checking and variable initialization. Version `mult-octave` is an implementation of array multiplication in Octave/MATLAB. First two arrays of size  $100 \times 100$  are filled with random numbers and every version is called to do the multiplication. Then two arrays of size  $300 \times 300$  are filled with random numbers and the same process is repeated. Results are shown in Table 1. As you can see `tym` versions are far more efficient than Octave/MATLAB version. This is because of dynamic nature of Octave/MATLAB as explained in Section 1.

Among versions which are implemented in `tym`, `mult-int-check` is the slowest one. The main reason is the bound checking that it does for every array look-up and array assignment.

It might be expected that `mult-int` be faster than `mult-real` since operations on integers are faster than operations on floating-points. However, `realArrays` are translated into `NDArrays` which are arrays of doubles (i.e `Array<double>`) but `intArrays` are translated into `int32NDArrays` which are arrays of `octave_int32`. `octave_int32` is a wrapper for integer that is provided by Octave to comply with MATLAB’s saturation semantic. Several operators are overloaded for `octave_int32` but to achieve efficiency `tymc` generates code that does not use these overloaded operators. For example `z(i, j) = z(i, j) + x(i, k)*y(k, j)` would be translated into

```
z(i, j) = z(i, j).value() + x(i, k).value() * y(k, j).value()
```

in case that `x` and `y` and `z` are `intArrays`. In this way, multiplication and addition are done on plain integers (and not wrapped `octave_int32s`) but this also incurs an additional overhead of calling `value` method of `octave_int32` objects whenever it is used in an expression. That might be the major reason for `mult-int` being slower than `mult-real`.

## Acknowledgement

We also have to credit another work `ompc` [3] which is a MATLAB to python compiler. We have used many of their grammar rules and routines in our work.

## References

- [1] Octave. <http://www.gnu.org/software/octave/>.
- [2] PLY (Python Lex-Yacc). <http://www.dabeaz.com/ply/>.

- [3] Peter Jurica and Cees van Leeuwen. OMPC: an open-source MATLAB-to-Python compiler. *Frontiers in Neuroinformatics*, February 2009.
- [4] Dag Sverre Seljebotn. Fast numerical computations with Cython. In *Proceeding 8th Python in Science conference*, 2009.